DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models

DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models

Amr Keleg and Walid Magdy
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
a.keleg@sms.ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

A few benchmarking datasets have been released to evaluate the factual knowledge of pretrained language models. These benchmarks (e.g., LAMA, and ParaRel) are mainly developed in English and later are translated to form new multilingual versions (e.g., mLAMA, and mParaRel). Results on these multilingual benchmarks suggest that using English prompts to recall the facts from multilingual models usually yields significantly better and more consistent performance than using non-English prompts. Our analysis shows that mLAMA is biased toward facts from Western countries, which might affect the fairness of probing models. We propose a new framework for curating factual triples from Wikidata that are culturally diverse. A new benchmark **DLAMA-v1** is built of factual triples from three pairs of contrasting cultures having a total of 78,259 triples from 20 relation predicates. The three pairs comprise facts representing the (Arab and Western), (Asian and Western), and (South American and Western) countries respectively. Having a more balanced benchmark (DLAMA-v1) supports that mBERT performs better on Western facts than non-Western ones, while monolingual Arabic, English, and Korean models tend to perform better on their culturally proximate facts. Moreover, both monolingual and multilingual models tend to make a prediction that is culturally or geographically relevant to the correct label, even if the prediction is wrong.

1 Introduction

Transfer learning paradigms such as fine-tuning, few-shot learning, and zero-shot learning rely on pretrained language models (PLMs), that require having large compilations of raw data (Devlin et al. 2019; Brown et al. 2020; Chowdhery et al. 2022; Scao et al. 2022). These PLMs showed some ability to model different linguistic phenomena (Goldberg 2019; Jawahar et al. 2019) in addition to memorizing facts related to real-world knowledge. While there is a drive to have multilingual models, English is still the language that is better supported due to the abundance of large English raw corpora, diverse datasets, and benchmarks. Moreover, monolingual non-English PLMs are still being pretrained for other high-resource languages. As a way to probe the non-English and multilingual PLMs, researchers tend to translate English benchmarks into other languages, which might degrade the quality of the samples especially if the translation is performed automatically. While translating English benchmarks saves the time and money needed to build new language-specific benchmarks, it might introduce unintended biases or artifacts into the benchmarks.

LAMA (Petroni et al., 2019) and ParaRel (Elazar et al., 2021) are two benchmarks developed to quantify the factual knowledge of the English PLMs. They used a setup in which a language model is said to know a specific fact if it can predict the right object for a prompt in a fill-the-gap setup (e.g., For the prompt **"The capital of England is [MASK]"**, the model needs to fill the masked gap with **"London"**). Multilingual versions of these benchmarks namely: mLAMA (Kassner et al., 2021), and mParaRel (Fierro and Søgaard, 2022) were released to evaluate the performance of multilingual PLMs by translating LAMA and ParaRel into 53 and 46 languages respectively. The subjects and objects of the triples within these benchmarks were translated using their multilingual labels on Wikidata, while the templates were automatically translated from the English ones used in the original benchmarks. These templates transform triples into textual natural language prompts for probing the models. X-FACTR is another benchmark sharing the same setup, and is built for 23 different languages (Jiang et al., 2020). All three benchmarks sample factual triples in the form of (subject, relation predicate, object) from T-REx, a dump of Wikidata triples aligned to abstracts extracted

Findings of the Association for Computational Linguistics: ACL 2023, pages 6245–6266
July 9-14, 2023

from the English Wikipedia (Elsahar et al., 2018). The way T-REx is constructed might make it more representative of the facts related to Western cultures, which might introduce an unnoticed bias to the benchmarks based on it. We hypothesize that having a fair representation of the different cultures within a benchmark is vital for fairly probing models pretrained for multiple languages. The main contributions of our paper can be summarized as follows:

1. Investigating the impact of sampling mLAMA triples from T-REx on the distribution of the objects within the relation predicates.
2. Proposing DiverseLAMA (DLAMA), a methodology for curating culturally diverse facts for probing the factual knowledge of PLMs, and building 3 sets of facts from pairs of contrasting cultures representing the (Arab-West), (Asia-West), and (South America-West) cultures, to form DLAMA-v1[1].
3. Showing the impact of having a less skewed benchmark DLAMA-v1 on the performance of mBERT and monolingual Arabic, English, Korean, and Spanish BERT models.
4. Demonstrating the importance of having contrasting sets of facts in diagnosing the behavior of the PLMs for different prompts.

2 Related Work

Petroni et al. (2019) investigated the possibility of using PLMs as potential sources of knowledge, which can later substitute manually curated knowledge graphs. To this end, they created LAMA (LAnguage Model Analysis), a dataset of 34,000 relation triples representing facts from 41 different Wikidata relation predicates. These facts are extracted from a larger dataset called T-REx that contains 11 million relation triples, acquired from a large Wikidata dump of triples, that were automatically aligned to English Wikipedia abstracts (Elsahar et al., 2018). Manual English templates were written to transform the triples into prompts to probe the model's factual knowledge. The triples were limited to the ones whose objects are tokenized into a single subtoken.

Kassner et al. (2021) constructed a multilingual version of LAMA (mLAMA) having 53 different languages. They handled the limitation of using single-subtoken objects by computing the probability of a multi-subtoken object as the geometric mean of the subtokens' probabilities. They concluded that the performance of mBERT when probed with prompts written in 32 languages is significantly lower than mBERT's performance when probed with English prompts. Moreover, they observed insignificant performance improvement for German, Hindi, and Japanese when their corresponding templates were manually corrected.

Similarly, Jiang et al. (2020) created X-FACTR by sampling relation triples from T-REx for 46 different Wikidata predicates. The multilingual Wikidata labels were used to translate the subjects and objects of the triples. They compared multiple decoding methods. Moreover, they employed different templates to generate prompts having the correct number/gender agreement with the subjects of the triples. English prompts still outperformed prompts written in 22 other languages.

ParaRel and its multilingual version mParaRel are benchmarks created by sampling triples from T-REx for 38 relation predicates (Elazar et al. 2021; Fierro and Søgaard 2022). Their aim is to measure the consistency of the model in making the same prediction for different paraphrases of the same template. Results on both benchmarks showed that the multilingual mBERT and XLM-R models are less consistent than the monolingual English BERT model, especially when these multilingual models are prompted with non-English inputs.

From a model diagnostics perspective, Cao et al. (2021) found that English PLMs might be biased to making specific predictions based on a predicate's template irrespective of the subjects used to populate this template. Thereafter, Elazar et al. (2023) designed a causal framework for modeling multiple co-occurrence statistics that might cause English PLMs to achieve high scores on some of LAMA's predicates.

We focus on why a non-English PLM might fail to recall facts and hypothesize the following possible reasons:

1. The quality of the template might degrade after automatically translating it from English.
2. Non-English or multilingual PLM are generally pretrained on a lesser amount of non-English data and thus might be less capable of recalling facts efficiently.
3. Translating the underlying facts of a benchmark, initially designed to probe English PLMs, might cause a representational bias.

[1]The DLAMA-v1 benchmark and the codebase can be reached through: https://github.com/AMR-KELEG/DLAMA

While the first two factors are studied in the literature, we believe that the third factor is a major quality issue that previous work has overlooked. Randomly sampling the triples from T-REx might introduce a representation bias toward Western cultures, since only facts aligned to English Wikipedia abstracts are considered. We investigate the presence of such bias (§3). Moreover, we empirically demonstrate how better model diagnostics can be performed when the benchmark is formed using two diverse and contrasting sets of facts (§5).

3 Cultural Bias in mLAMA

Probing PLMs using prompts is an analysis tool attempting to understand how they behave. A biased probing benchmark might be deceiving, as both a good-performing model and a model sharing the same bias found in the benchmark would achieve good performance. In this section, we investigate if the facts within mLAMA might be biased toward Western cultures, which can affect the reliability of the performance scores achieved by PLMs when probed using mLAMA.

3.1 Quantifying the Cultural Bias

As a proxy for measuring the skewness of the triples of T-REx, LAMA, and X-FACTR toward Western cultures, 26 relation predicates are selected that have a person's name or a place as their subject or object. Moreover, 21 Western countries are identified as representative of Western cultures from Western European and South Western European countries[2]: Andorra, Austria, Belgium, France, Germany, Ireland, Italy, Liechtenstein, Luxembourg, Monaco, Netherlands, Portugal, San Marino, Spain, Switzerland, the United Kingdom, in addition to Canada, the United States of America, Australia, and New Zealand. For each relation predicate out of the 26, triples with a subject or object that either has a country of citizenship or is located in one of the 21 Western countries are counted.

63.6% of the triples within the LAMA benchmark are related to these Western countries compared to 62.7% for X-FACTR, and 57.1% for T-REx (from which LAMA and X-FACTR are sampled)[3]. This highlights the issue that aligning Wikidata triples to English Wikipedia abstracts in T-REx would skew them toward Western countries, impacting both LAMA and X-FACTR.

3.2 Qualitative Analysis of the Bias and its Impact

Kassner et al. (2021) used mLAMA to probe mBERT using prompts in 41 languages. We find that all the languages in which prompts achieve the highest performance[4] use the Latin script, while the ones with the least performance[5] use other scripts. This might be attributed to the model's ability to share cross-lingual representations for common named entities for languages using the Latin script, which allows for cross-lingual knowledge sharing. Moreover, it is known that more than 78% of mBERT's vocabulary is Latin subwords[6].

However, there are still some relation predicates for which a non-Latin scripted language outperforms a Latin-scripted one. The P140[7] (religion or worldview) predicate is a clear example of these predicates. An example triple for the P140 predicate is: **(Edward I of England, religion or worldview [P140], Christianity)**. mBERT has higher performance for Arabic (23.1%), Azerbaijani (8.1%), Korean (30.1%), Georgian (35.1%), Thai (13.4%), Tamil (4.0%), Russian (54.6%), and Japanese (30.0%) than for English (1.5%). Looking at the objects for the English triples within mLAMA, we find that 53.7% of the triples have *Islam* as their object.

While the objects for the P140 predicate should be religions, we find that only seven triples have incorrect inflected forms of *Muslim*, *Christian*, and *Hindu* instead of *Islam*, *Christianity*, and *Hinduism*. Further investigation reveals that the English template used to transform the triples into prompts is (*[X] is affiliated with the [Y] religion .*) which would suit retrieving these infrequent inflected labels than the frequent labels. Therefore, most predictions for the English prompts are considered incorrect justifying the low performance achieved for English. To overcome penalizing these predictions, we mapped the model's predictions and the objects' labels such that for instance *Christian* and *Christianity* are both considered to represent the same prediction *Christianity*, and similarly for *Hinduism* and *Islam*.

[2]According to EuroVoc: https://eur-lex.europa.eu/browse/eurovoc.html?params=72,7206#arrow_912

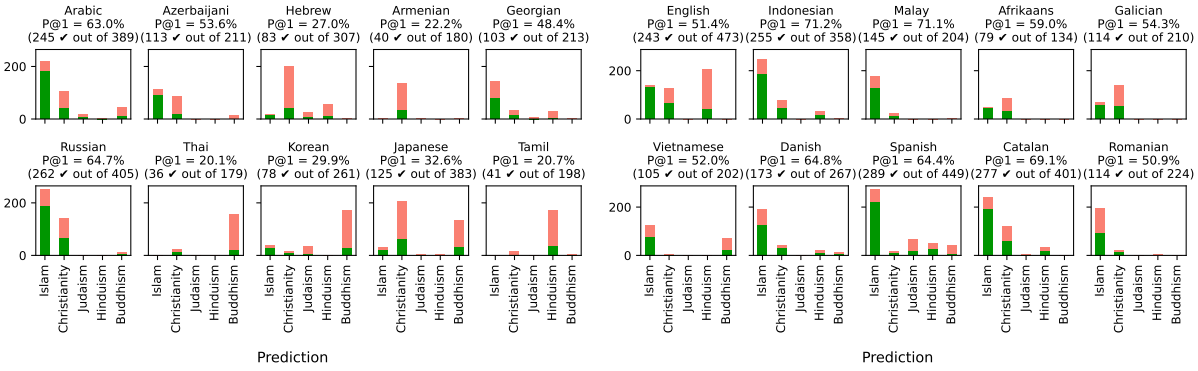
[3]Full percentages for each predicate are listed in Table A1.

[4]English, Indonesian, Malay, Afrikaans, Galician, Vietnamese, Danish, Spanish, Catalan, Cebuano, Romanian.

[5]Russian, Azerbaijani, Hebrew, Arabic, Korean, Armenian, Georgian, Tamil, Thai, Japanese.

[6]http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html

[7]Wikidata predicates' identifiers format is $P[0-9]+$.



(a) Languages with the least overall performance on mLAMA. (b) Languages with the highest overall performance on mLAMA.

Figure 1: The distribution of the predictions of mBERT for the P140 predicate (religion or worldview) for prompts in 20 different languages after merging similar objects' predictions (e.g., Muslim and Islam). The green portion of the bar represents the triples for which the prediction is correct, while the red portion represents the triples for which the prediction is wrong. The **P@1 (Precision at first rank)** metric is the percentage of triples for which the model's first prediction for a triple's subject matches the triple's object.
Note: **P@1** scores are not directly comparable since the number of triples in mLAMA differs between languages.

Figure 1 shows the distribution of mBERT's predictions for the P140 triples for prompts in 20 different languages after unifying the labels. We observe that: **(1)** For some languages, the predictions are skewed toward a specific wrong label that is culturally related to these languages. For example, the mode of the predictions of prompts in Armenian, Thai, Korean, and Tamil is Christianity, Buddhism, Buddhism, and Hinduism respectively. **(2)** Arabic, and Russian prompts tend to yield high performance. The same holds for Indonesian and Malay which achieve similar performance with less skewness in the predictions. Since the label distribution for this predicate within mLAMA is skewed toward a specific label *Islam*, one can not confidently conclude whether the model is choosing the right answer for having some knowledge of the facts or for making a biased guess that luckily coincides with the right label. While these findings signify the possibility that mLAMA is biased for the P140 predicate, it on the other hand might hint that mLAMA is also biased toward Western cultures for most of the remaining predicates. For instance, the P103 (Native Language) predicate in mLAMA has *French* as the correct label for 60.14% of the triples.

4 Building DLAMA

Our methodology aims at building a culturally diverse benchmark, which would allow for a fairer estimation of a model's capability of memorizing facts. Within DLAMA, query parameters form underlying SPARQL queries that are used to retrieve Wikidata triples as demonstrated in Figure 2.

To operationalize the concept of cultures, we use countries as a proxy for the cultures of interest. For instance, countries that are members of the Arab League are considered representatives of Arab cultures. Conversely, Western countries mentioned in §3.1 represent Western cultures. Furthermore, China, Indonesia, Japan, Malaysia, Mongolia, Myanmar, North Korea, Philippines, Singapore, South Korea, Taiwan, Thailand, and Vietnam are 13 countries from East Asia, and Southeast Asia[8] representing Asian cultures, while Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela represent South American cultures.

For predicates in which the subject is a person, we add a filter to the SPARQL query which limits the country of citizenship of the person to a specific set of countries (i.e., a specific culture). For predicates in which the subject is a place, we limit the values of the places to those located in a country within the predefined set of countries related to the target culture.

We implemented a Python interface to simplify the process of querying Wikidata triples. Currently, 20 relation predicates are supported. The user-friendly interface allows the addition of new relation predicates and filters, which we hope would encourage contributions to DLAMA.

[8] Based on the UN stats classification: https://unstats.un.org/unsd/methodology/m49/

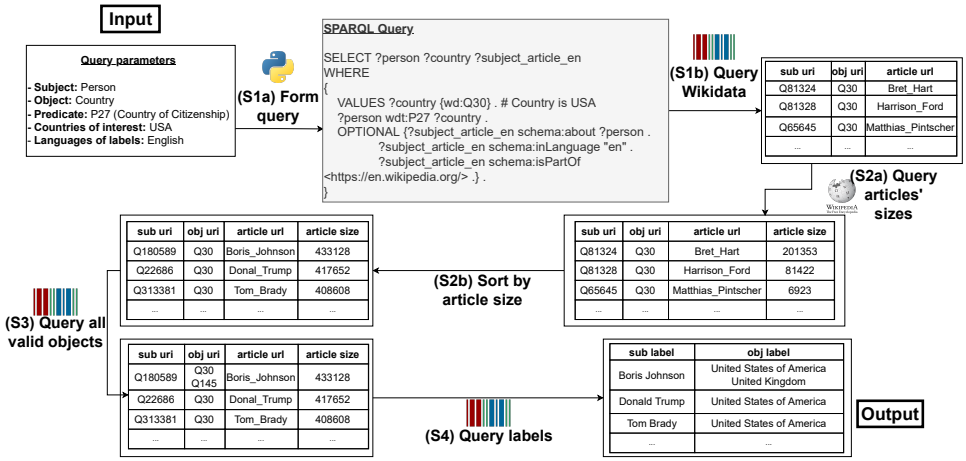


Figure 2: A demonstration of DLAMA's querying framework for the predicate P27 (Country of Citizenship).

4.1 Methodology of Querying Triples for a Specific Predicate

Step #1 - Getting an exhaustive list of triples for a Wikidata predicate: A set of parameters need to be specified through the Python interface to generate an underlying SPARQL query. These parameters are **(1)** an entity label for the subject, and an entity label for the object[9], **(2)** a set of countries representing specific cultures, **(3)** a Wikidata predicate relating the object to the subject, **(4)** a list of Wikipedia sites that are expected to contain facts related to each specified country, and **(5)** a list of languages for which the parallel labels of the subjects and the objects are acquired and later used to populate the multilingual probing templates. In addition to querying the Wikidata Unique Reference Identifiers (URIs) of the subjects and the objects, the Unique Reference Links (URLs) of the Wikipedia articles linked to the subjects are queried as optional fields.

Step #2 - Sorting the list of retrieved triples by their validity: Facts on Wikidata are crowdsourced, and contributors are encouraged to add references to the facts they modify. However, lots of the facts on Wikidata still have missing references. Therefore, we use the length of the Wikipedia article corresponding to the triple's subject as a proxy for the validity of the triple. The fact that contributors and editors spent time writing a long Wikipedia article implies that a group of people finds the article important. Therefore they will be keen on making sure the information there is factually sound (Bruckman, 2022). We believe that using the size of the article rather than other metrics such as the number of visits to the Wikipedia article, allows facts related to underrepresented groups on Wikipedia to still be ranked high, thus making the top-ranked facts more diverse and inclusive. We sort the retrieved triples by the size (in bytes) of the Wikipedia article linked to their subjects. In case a subject has articles on multiple Wikipedia sites, the size of the largest article is used. DLAMA also allows sorting the triples by the total number of edits (revisions) of their subjects' respective articles.

Step #3 - Querying all possible objects for each subject: Since a subject might be linked to multiple objects for the same relation predicate, another query is executed in order to ensure that all these objects are retrieved. For instance, a person might be a citizen of an Arab country in addition to another non-Arab country. This step ensures that the non-Arab country is still considered as a valid country of citizenship for the person, even if the initial query restricted the countries to Arab ones only. While previous benchmarks limited the object for each triple to a single value, we believe it is fairer to allow multiple valid labels instead of randomly picking one label out of the valid ones.

Step #4 - Querying the labels for the triples: Till this stage, the subjects and objects are represented by their Wikidata URIs. The Wikidata labels of all the subjects and objects need to be fetched for the languages of interest. Relation triples having missing subject or object labels in any of the languages specified are discarded in order to ensure that the triples are the same for all the languages.

[9]**The used entity labels are** City, Continent, Country, Genre, Instrument, Language, Occupation, Original Network, Person, Piece of Work, Place, Record Label.

Step #5 (optional) - Handling overlapping objects: The degree of granularity of the objects for Wikidata's relation predicates differs even among triples of the same predicate (e.g.: The official language of `Australia` is set to `English` while that of `The United States of America` is set to `American English` which is a subclass of `English`). To avoid penalizing models for picking an object that is a superclass of the correct object, a graph is built, modeling the hierarchical relations between all the objects of the sampled triples of a relation predicate. The graph is later used to augment the valid objects with their superclasses as detailed in §B of the Appendix.

4.2 The DLAMA-v1 Benchmark

We used the above method to build three sets of facts as part of DLAMA-v1 to assess the performance of PLMs on recalling facts related to 21 Western countries as compared to the 22 Arab, 13 Asian countries, and 12 South American countries[10]. The sets provide examples of how the framework can be used to compile facts from pairs of contrasting cultures. We hope the community will use the framework to introduce new pairs representing other countries and cultures. A maximum of 1000 triples from each predicate out of the 20 supported ones are independently queried for each set of countries within each pair. This ensures that the queried triples are balanced across the two sets of countries within the pair.

In total, the (Arab-West) pair comprises 24535 triples with labels in Arabic and English, as compared to 27076 triples with labels in Korean, and English for the (Asia-West) pair, and 26657 triples with labels in Spanish, and English for the (South America-West) pair. Figure 3 shows an example of a triple of DLAMA-v1's (Arab-West) set. The underlying triples belonging to the Western cultures in the 3 sets are not identical. Triples in a set are discarded if their subjects or objects do not have labels in the languages.

Regarding the languages of the labels, Arabic and Korean are chosen as they are two of the least-performing languages on mLAMA. It is expected that facts related to Arab and East Asian/South East Asian countries are relevant to Arabic and Korean PLMs respectively, and would be contrasting to Western facts. Additionally, both languages have non-Latin scripts, use white spaces to separate tokens, and have an inventory of monolingual PLMs. On the other hand, the (South America-West) pair is a trickier case since most South American countries use Spanish as their official language. One can argue that sharing the same language with Spain introduces commonalities between the South-American countries and the Western ones.

[10]Refer to §3.1 and §4 for the list of countries.

- **Prompt**: Egypt is located in ...
- **Subject**: {Egypt}
- **Set of correct objects**: {Africa, Asia}
- **Set of objects of the predicate to be ranked**: {Africa, Asia, Europe, Insular Oceania, North America}

Figure 3: An example of a prompt created using a relation triple of DLAMA from the P30 (continent) relation predicate for the Arab-Western pair.

Overlap between DLAMA-v1 and T-REx: For the three culture sets, we measured the percentage of triples found in T-REx. 17.92% of Arab-related facts are in T-REx compared to 39.85% of Western-related ones in the (Arab-Western) pair. Moreover, 22.64% of Asian-related facts are found in T-REx compared to 44.43% of Western-related ones in the (Asia-Western) pair. Lastly, the overlap percentages for the (South America-West) pair are 17.68% and 32.22% respectively. These values demonstrate that T-REx has less coverage of the Arab, Asian, and South American factual triples than its coverage of Western triples. Moreover, the fact that T-REx is tuned for higher precision means that its recall is affected and a lot of the Western facts expected to be found in English Wikipedia abstracts are discarded. Conversely, DLAMA-v1 is a less skewed benchmark across different cultures.

5 Probing PLMs via DLAMA-v1

5.1 Experimental Setup

We follow mLAMA's probing setup to evaluate the PLMs' factual knowledge. For each relation predicate [PREDICATE], the set {OBJECTS} of unique objects of the triples is first compiled. Then, for each relation triple within the [PREDICATE], the PLM is asked to assign a score for each object within {OBJECTS} by computing the probability of having this object replacing the masked tokens. This setup asks the model to choose the correct answer out of a set of possible choices, instead of decoding the answer as a generation task. The templates used in DLAMA to convert triples into natural language prompts are adapted from mLAMA and listed in Table F9 of the Appendix.

Prompt Lang.	Model name	$P_{@1}$ Arab N=10946	$P_{@1}$ West N=13589	$P_{@1}$ DLAMA N=24535	$P_{@1}$ mLAMA N=17128
Arabic	mBERT-base	13.7	**15.1***	14.5	**15.2†**
	arBERT	**33.6***	23.0	**27.7†**	24.4
English	mBERT-base	21.2	**37.7***	30.3	**33.9†**
	BERT-base	27.5	**31.3***	29.6	**37.9†**

(a) DLAMA-v1 (Arab-West)

Prompt Lang.	Model name	$P_{@1}$ Asia N=13479	$P_{@1}$ West N=13588	$P_{@1}$ DLAMA N=27067	$P_{@1}$ mLAMA N=14217
Korean	mBERT-base	16.4	**28.5***	**22.5†**	15.7
	KyKim	**22.1***	19.5	**20.8†**	13.4
English	mBERT-base	33.0	**39.9***	**36.4†**	35.1
	BERT-base	**38.3***	31.9	35.1	**39.0†**

(b) DLAMA-v1 (Asia-West)

Prompt Lang.	Model name	$P_{@1}$ S. America N=13071	$P_{@1}$ West N=13586	$P_{@1}$ DLAMA N=26657	$P_{@1}$ mLAMA N=28168
Spanish	mBERT-base	25.4	**33.8***	29.7	**30.5†**
	BETO	16.0	**26.5***	21.4	**22.7†**
English	mBERT-base	27.0	**37.6***	32.4	**33.9†**
	BERT-base	26.9	**31.3***	29.2	**37.1†**

(c) DLAMA-v1 (South America-West)

Table 1: Performance of mBERT, and monolingual Arabic (arBERT), Korean (KyKim), Spanish (BETO), and English (BERT-base) language models on the three sets of facts of DLAMA-v1. *: the set of cultures on which a model performs better, †: the benchmark on which the model achieves higher $P_{@1}$ score.

Models: We evaluated the cased multilingual BERT-base, and the cased English BERT-base using all the sets of facts of DLAMA-v1. Moreover, a monolingual Arabic BERT-base model **arBERT** (Abdul-Mageed et al., 2021), a monolingual Korean BERT-base model **KyKim BERT-base** (Kim, 2020), and a monolingual cased Spanish BERT-base model **BETO** (Cañete et al., 2020) are evaluated using the (Arab-West), the (Asia-West), and the (South America-West) pairs respectively. We focus on BERT models to compare our results to those previously reported on mLAMA.

5.2 Aggregated Results

Precision at the first rank (P@1) is the metric used to evaluate the performance of the models. P@1 is the percentage of triples for which the first prediction of the model matches one of the objects for this triple. In order to quantify the diversity of the objects of a relation predicate for each culture, an entropy score is computed. For each triple of a relation predicate, only the most frequent object among the list of valid objects is considered. The entropy score is computed as $Entropy(\{objs\}) = \sum_{o \in \{objs\}} -p_o * \log(p_o)$; where p_o is the probability of object o across the set of objects $\{objs\}$. The higher the entropy of the objects is, the more diverse the objects are, and thus the harder the predicate would be for a model to randomly achieve high (P@1) scores.

Looking at the performance of models on DLAMA indicated in Table 1, **(1)** we find how the facts' relevance to the probed model's language affects the results. For instance, arBERT and KyKim perform better on non-Western facts than on Western ones. Conversely, the English BERT-base model performs better on Western facts for the (Arab-West) pair. The same observation tends to hold for individual predicates as shown in Table 2. **(2)** Moreover, arBERT and KyKim achieve lower performance on mLAMA than their performance on DLAMA-v1, while the English BERT-base and BETO models achieve higher $P_{@1}$ scores on mLAMA than on DLAMA-v1. This is expected given the bias mLAMA has toward facts from Western cultures.

5.3 Revisiting the Language bias of PLMs

Kassner et al. (2021) showed that for prompts in English, German, Dutch, and Italian, mBERT is biased toward predicting the language or the country name related to the language of the prompts (e.g., Filling the masked object with *Italy* if the prompt's language is *Italian*). This phenomenon is not a bias if most of the triples in the underlying subset of mLAMA for a language are also biased toward the same label. For DLAMA, looking at the P@1 scores in Table 2 in addition to checking the most common predictions of arBERT and the cased BERT-base models in Table 3 provides a better diagnostic tool for analyzing the models' behavior[11]. For the P364 predicate, the models perform better on their culturally proximate triples. This can be attributed to the Language bias phenomenon which is indicated by arBERT predicting *Arabic* for 30.8% of Western facts, while BERT-base predicting *English* for 44.6% of Arab facts. On the other hand, both models achieve high P@1 scores for P17 and P103. Even when the models make wrong predictions for triples of these predicates, the predictions can be considered to be educated guesses, as they are still relevant to the culture to which the triples belong. Lastly, the models perform poorly on P495 for being biased toward specific objects irrespective of the culture of the triples (*Japan* for BERT-base, *Germany* and *France* for arBERT). These three patterns can be noticed thanks to having a contrastive set of facts representing two different cultures.

[11] A similar analysis for the other two sets of contrasting cultures can be found in §E of the Appendix.

Relation	# facts (entropy)		Arabic prompts P@1		English prompts P@1	
	Arab	West	Arab	West	Arab	West
P17 (Country)	1000 (3.9)	1000 (2.8)	**49.9**	47.4	**52.2**	45.6
P19 (Place of birth)	1000 (3.9)	1000 (2.6)	**33.7**	22.3	**10.1**	8.8
P20 (Place of death)	1000 (3.8)	1000 (2.7)	21.3	**22.7**	14.2	**17.2**
P27 (Country of citizenship)	1000 (3.8)	1000 (2.4)	**38.1**	27.9	4.1	**17.5**
P30 (Continent)	22 (1.0)	19 (1.0)	**45.5**	26.3	**86.4**	84.2
P36 (Capital)	22 (4.5)	19 (4.2)	**95.5**	78.9	36.4	**84.2**
P37 (Official language)	22 (0.0)	19 (2.5)	**90.9**	84.2	95.5	**100.0**
P47 (Shares border with)	22 (2.5)	19 (2.7)	**27.3**	15.8	68.2	**78.9**
P103 (Native language)	1000 (1.0)	1000 (1.7)	61.8	**72.8**	67.7	**74.4**
P106 (Occupation)	1000 (2.3)	1000 (2.0)	**3.7**	3.3	4.8	**14.3**
P136 (Genre)	452 (2.7)	1000 (2.6)	6.6	**24.3**	4.0	**7.6**
P190 (Sister city)	67 (4.9)	468 (7.3)	0.0	**2.6**	**6.0**	2.8
P264 (Record label)	166 (3.0)	1000 (5.2)	0.0	**0.3**	4.2	**7.5**
P364 (Original language of work)	1000 (0.6)	1000 (0.4)	**61.2**	48.5	36.1	**88.9**
P449 (Original network)	127 (4.5)	1000 (5.3)	**0.8**	0.4	0.0	**10.8**
P495 (Country of origin)	1000 (3.1)	1000 (1.3)	**18.6**	8.7	**14.7**	5.5
P530 (Diplomatic relation)	22 (0.0)	19 (0.0)	22.7	**42.1**	31.8	**68.4**
P1303 (Instrument)	1000 (0.9)	1000 (1.1)	**0.3**	0.2	1.9	**27.7**
P1376 (Capital of)	24 (4.3)	26 (4.0)	**91.7**	84.6	**79.2**	76.9
P1412 (Languages spoken or published)	1000 (0.8)	1000 (1.5)	**67.4**	26.1	83.4	**88.7**
Aggregated statistics	10946 (2.6)	13589 (2.7)	**33.6**	23.0	27.5	**31.3**

Table 2: Detailed P@1 scores of arBERT (Arabic prompts) and cased BERT-base (English prompts) on the DLAMA-v1 (Arab-West) set. **Note:** $\#\ facts$ is the number of facts for each culture within the benchmark, while $(entropy)$ is the entropy of the objects for the facts of each culture.

5.4 Pilot Evaluation for a Large Language Model

Given the success of large language models (LLMs) (Brown et al., 2020; Scao et al., 2022), we evaluated the performance of the GPT3.5-turbo model on tuples from the P30, P36, P37, P47, P103, P530, and P1376 predicates of DLAMA-v1 (Arab-West). To probe the model, the Arabic and English templates for these predicates were mapped into questions listed in Table F10. While the model is instructed to only respond with an entity, it sometimes provides a full sentence. Consequently, we consider the model's response to a question to be correct if one of the valid objects of the tuple used to populate the question is a substring of the response. GPT3.5's probing setup is harder than BERT's setup in which an answer is chosen from a set of unique objects for the predicate. Nevertheless, GPT3.5 achieves superior performance compared to the monolingual BERT models as per Table D3. However, GPT3.5 seems to be hallucinating for a lot of the tuples within the P190 (Sister City) predicate (e.g.: **The twin city of Nice is Naples.**). Such issues might be unnoticed unless benchmarks like DLAMA are used to systematically evaluate the LLMs.

6 Conclusion

Previous work suggested that English prompts are more capable of recalling facts from multilingual pretrained language models. We show that the facts within the underlying probing benchmark (mLAMA) are skewed toward Western countries, which makes them more relevant to English. Hence, we propose a new framework (DLAMA) that permits the curation of culturally diverse facts directly from Wikidata. Three new sets of facts are released as part of the DLAMA-v1 benchmark containing factual triples representing 20 relation predicates comprising facts from (Arab-Western), (Asian-Western), and (South American-Western) countries, with a more balanced representation between the countries within each pair. The results of probing PLMs on the DLAMA-v1 support that mBERT has a better performance recalling Western facts than non-Western ones irrespective of the prompt's language. Monolingual Arabic and Korean models on the other hand perform better on culturally proximate facts. We believe the probing results are more trustable and fairer when the underlying benchmark is less skewed toward specific countries, languages, or cultures. Moreover, we find that even when the model's prediction does not match any of the correct labels, the model might be making an educated guess relevant to the culture of the underlying facts. This finding augments previous experiments which showed that models tend to have a language bias, by which a model tends to overgenerate a specific prediction for each prompting language irrespective of the triple's subject used to fill in the prompt. Finally, our framework is open-sourced for the community to contribute new pairs to the DLAMA benchmark in the future.

Relation predicate	Common correct predictions (% of predictions)	Common wrong predictions (% of predictions)
Probing arBERT with Arabic prompts populated with Arab facts		
P17: [X] is located in [Y] .	Egypt (8.4%) Algeria (7.5%) Morocco (5.6%)	Morocco (10.5%) Turkey (7.1%) Tunisia (6.7%)
P19: [X] was born in [Y] .	Egypt (9.0%) Algeria (4.2%) Morocco (3.9%)	Algeria (18.8%) Tunisia (7.2%) Morocco (5.9%)
P20: [X] died in [Y] .	Egypt (9.8%) Baghdad (2.0%) Tunisia (1.1%)	Paris (28.7%) Egypt (8.8%) Tunisia (5.7%)
P27: [X] is [Y] citizen .	Saudi Arabia (4.5%) Morocco (4.2%) Egypt (3.8%)	State of Palestine (15.2%) Republic of Egypt (9.3%) Iraqi Republic (5.2%)
P495: [X] was created in [Y] .	Egypt (6.4%) Morocco (3.0%) France (2.9%)	France (26.9%) Germany (19.5%) Morocco (10.2%)
P103: The native language of [X] is [Y] .	Arabic (59.7%) French (1.1%) English (0.4%)	English (8.8%) Arabic (8.6%) Shilha (8.4%)
P364: The original language of [X] is [Y] .	Arabic (58.9%) French (1.1%) English (0.9%)	Shilha (11.7%) English (8.5%) French (7.6%)
P1412: [X] used to communicate in [Y] .	Arabic (62.5%) French (4.7%) Spanish (0.1%)	Syrian Arabic (15.1%) Arabic (9.1%) French (5.0%)
Probing arBERT with Arabic prompts populated with Western facts		
P17: [X] is located in [Y] .	France (13.5%) United States of America (9.3%) Spain (9.1%)	Germany (8.3%) South Africa (7.5%) France (6.6%)
P19: [X] was born in [Y] .	Germany (7.4%) Italy (4.6%) New York City (3.5%)	New York City (26.0%) Germany (18.0%) Italy (10.8%)
P20: [X] died in [Y] .	Paris (9.0%) Germany (3.5%) Italy (3.4%)	Paris (33.4%) New York City (19.7%) London (7.6%)
P27: [X] is [Y] citizen .	France (8.7%) Germany (6.5%) United States of America (6.3%)	Germany (12.9%) French protectorate of Tunisia (11.0%) Republic of Ireland (6.8%)
P495: [X] was created in [Y] .	United States of America (3.9%) France (2.4%) Germany (1.7%)	Germany (44.2%) France (23.4%) Algeria (4.5%)
P103: The native language of [X] is [Y] .	English (53.1%) French (12.8%) German (3.2%)	French (7.7%) English (4.2%) Spanish (3.9%)
P364: The original language of [X] is [Y] .	English (47.2%) French (0.7%) German (0.2%)	Arabic (30.8%) French (5.7%) Shilha (5.1%)
P1412: [X] used to communicate in [Y] .	French (12.6%) German (7.7%) Spanish (3.4%)	Arabic (55.9%) German (8.3%) French (3.4%)
Relation predicate	**Common correct predictions (% of predictions)**	**Common wrong predictions (% of predictions)**
Probing BERT-base with English prompts populated with Arab facts		
P17: [X] is located in [Y] .	Algeria (9.0%) Egypt (8.6%) Iraq (4.6%)	Bahrain (8.6%) Moscow (4.1%) Lebanon (3.8%)
P19: [X] was born in [Y] .	Cairo (3.8%) Baghdad (3.0%) Damascus (0.5%)	Baghdad (31.1%) Cairo (18.1%) Paris (6.4%)
P20: [X] died in [Y] .	Cairo (10.9%) Baghdad (2.0%) Egypt (0.7%)	Cairo (45.9%) Paris (19.2%) Baghdad (8.0%)
P27: [X] is [Y] citizen .	France (1.8%) Qatar (1.3%) Israel (0.4%)	Qatar (73.9%) Pakistan (8.8%) Israel (2.8%)
P495: [X] was created in [Y] .	Egypt (10.0%) Algeria (1.1%) Iraq (0.7%)	Japan (25.2%) India (12.2%) Egypt (9.2%)
P103: The native language of [X] is [Y] .	Arabic (66.2%) French (0.8%) English (0.4%)	Arabic (12.2%) Urdu (6.5%) Kurdish (4.3%)
P364: The original language of [X] is [Y] .	Arabic (30.7%) English (3.5%) French (1.6%)	English (44.6%) French (4.3%) Hindi (2.1%)
P1412: [X] used to communicate in [Y] .	Arabic (78.3%) English (2.8%) French (1.8%)	Arabic (8.5%) English (4.5%) Urdu (1.2%)
Probing BERT-base with English prompts populated with Western facts		
P17: [X] is located in [Y] .	France (15.7%) Spain (10.3%) Germany (8.2%)	Georgia (10.1%) Moscow (9.1%) Canada (6.0%)
P19: [X] was born in [Y] .	Paris (3.4%) Berlin (0.9%) London (0.7%)	Chicago (25.4%) London (22.1%) Paris (9.5%)
P20: [X] died in [Y] .	Paris (9.4%) London (2.7%) Rome (2.6%)	Paris (29.0%) London (22.9%) Rome (8.3%)
P27: [X] is [Y] citizen .	France (11.0%) Italy (2.5%) Canada (1.1%)	British America (44.5%) Austria (8.5%) Canada (6.8%)
P495: [X] was created in [Y] .	France (2.2%) Germany (1.4%) Japan (0.6%)	Japan (61.3%) England (10.1%) India (4.6%)
P103: The native language of [X] is [Y] .	English (50.8%) French (13.3%) German (3.3%)	Spanish (6.8%) German (3.6%) French (3.2%)
P364: The original language of [X] is [Y] .	English (85.4%) French (1.5%) German (0.9%)	Latin (2.1%) English (2.0%) French (1.9%)
P1412: [X] used to communicate in [Y] .	English (59.8%) French (13.1%) German (7.3%)	English (3.6%) Spanish (2.2%) Arabic (1.2%)

Table 3: The most common predictions for monolingual arBERT and BERT-base models when probed by DLAMA-v1 (Arab-West) with English and Arabic prompts respectively. Purple culturally related prediction, Blue bell culturally proximate prediction, Light Orange culturally proximate prediction to another culture, Orange culturally related prediction to the other culture. **Note:** The Arabic prompts/entities are translated for clarity.

Limitations

We acknowledge that the methodology used to build DLAMA-v1 still has limitations related to the information within its relation triples. While directly querying Wikidata as a dynamic source of facts provides the flexibility needed to acquire data that is relevant to different cultures (as opposed to using the static T-REx dump of triples), the diversity of the triples that are compiled depends on the availability of a diverse set of facts on Wikidata in the first place. For instance, the smaller number of relation triples related to Arab countries for the predicates (P136 - Genre), (P190 - Sister city), and (P449 - Original network) in DLAMA-v1 (Arab-West) demonstrates the difficulty of querying the exact number of facts for both cultures despite using exactly the same queries with the only difference being limiting the region to which the triples belong. Another limitation is the inability to enumerate valid and fine-grained subclasses of objects for specific subjects, if these fine-grained objects are not on Wikidata. Steps #3 and #5 of DLAMA explained in §4.1 ensure that a possible and more general object is still valid for a specific subject. However, inferring a more specified object from a generic one is impossible. For example, the fact that someone speaks “American English" implies that they speak English as well, but knowing that someone speaks “English" is not enough to speculate about their dialect (i.e.: “American English", “British English", etc.).

While the triples within DLAMA are sampled by picking the ones whose subjects have the largest Wikipedia articles’ sizes, the infeasibility of manually reviewing the large number of diverse facts within DLAMA-v1 makes it hard to claim that the facts are free of inaccuracies or missing information. More broadly, DLAMA supports relations predicates that are already part of mLAMA to fairly compare the results on DLAMA to those previously reported on mLAMA. Moreover, we make sure that the subjects and the objects of the relation triples are available in the different languages of interest. Having these constraints might imply that some culturally relevant facts might have been dropped out of DLAMA-v1 (e.g., Predicates that are not part of mLAMA, or triples having missing labels in one of the languages of interest).

Lastly, we used mLAMA’s probing setup in which the models rank a predefined set of objects for each prompt. Their prediction is correct if the top-ranked object is one of the valid labels for the corresponding relation triple used to populate the prompt. Therefore, a model’s performance is expected to be higher than that achieved by a generative setup in which the model is asked to generate the most probable completions for the masked tokens.

Ethics Statement

We believe that using a set of countries to represent cultures is just a proxy for acquiring a more diverse set of facts that are less skewed toward a specific culture. More specifically, using the terms Arab cultures, Western cultures, and Asian cultures simplifies the differences between the cultures within the countries that we have used to represent these macro-cultures. On the other hand, we still think that the differences between Asian cultures are less subtle than between them and Western cultures.

We also acknowledge that the accuracy and validity of some relation triples queried from Wikidata might be biased by the views of the people who added such information to Wikidata. This might be particularly vibrant for relation triples related to zones with political/ sectarian wars and conflicts.

Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics. Amr is grateful to Matthias Lindemann for recommending Wikidata, Aida Tarighat for the early discussions about the benchmark, Laurie Burchell, Bálint Gyevnár, and Shangmin Guo for reviewing the manual prompts, Coleman Haley for the multiple discussions about the figures, Anna Kapron-King and Gautier Dagan for proofreading the abstract, and lastly, Dilara Keküllüoğlu and Björn Ross for their valuable reviews of the paper’s final draft.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Amy S Bruckman. 2022. *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge*. Cambridge University Press.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arxiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2023. Measuring causal effects of data statistics on language model's 'factual' predictions.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *ArXiv*, abs/1901.05287.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Kiyoung Kim. 2020. Pretrained language models for korean. `https://github.com/kiyoungkim1/LMkor`.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

A Detailed Bias Values within the Factual Knowledge Benchmarks

Table A1 provides the fine-grained percentages for the distribution of the triples of T-REx, LAMA, and X-FACTR for 21 Western countries as compared to the rest of the world. For most of the relation predicates, triples related to one of the 21 Western countries represent more than 50% of the total triples. We find that this skewness is even larger for LAMA, and X-FACTR than for T-REx. Triples within LAMA are restricted to the ones whose objects are tokenized into a single subword by monolingual language models. This filtering might be responsible for the increased skewness of LAMA toward facts from Western countries.

B Augmenting the correct objects within DLAMA

For each relation predicate, a graph is used to model all the subclass-superclass relations between the objects of the queried triples. The edges within the graph are built using Wikidata's **P279 (subclass of)** predicate. All the possible subclass/superclass relations between the list of objects for each relation predicate are queried and then used to form the edges of the graph. Afterward, the list of objects for each subject is augmented by the list of all the possible ancestors (superclasses) of these objects (e.g., The official languages of `The United States of America` are now set to `American English` and `English` instead of just `American English`).

Similarly, we noticed that the level of specificity of places of birth (objects of P19) and places of death (objects of P20) varies between different tuples. Thus, we queried all the territorial entities in which the places of birth and death are located. For instance, `Paris Hilton` had the place of birth set to `{New York City}` while `Donald Trump` had the place of birth set to `{Jamaica Hospital Medical Center}`. After querying the higher administrative-territorial entities, the set of valid objects for both entities became `{New York City, New York, United States of America}` and `{Jamaica Hospital Medical Center, Queens, New York City, New York, United States of America}` respectively.

C Results on Raw Triples before the Last Optional Step

To demonstrate the impact of the last optional step within DLAMA, we evaluate the PLMs on the triples before augmenting their objects with valid overlapping ones (i.e.: before applying the optional Step #5 of the framework). It is clear that the performance of the models shown in Table C2 is worse than their performance on the augmented benchmark previously listed in Table 1.

D GPT3.5 performance on a subset of DLAMA (Arab-West)

As mentioned in §5.4, we used OpenAI's API to evaluate the performance of the GPT3.5-turbo model on six predicates of DLAMA-v1 (Arab-West). The accuracy scores of the model for these predicates are reported in Table D3. We plan to extend our evaluation to cover more predicates and include other LLMs.

E Model diagnostics using the (Asia-West) and (South America-West) sets

Contrasting KyKim BERT to English BERT-base: We replicate the analysis process done in §5.3 to investigate the behavior of KyKim BERT-base and the English BERT-base models using Tables E4 and E6. We find that the English BERT-base has the same patterns detailed before for P17, P103, P364, and P495. Moreover, since English BERT-base overgenerates *Japan* for the P495 predicate, its performance on the Asian part of DLAMA-v1 (Asia-West) is high. This once again shows the importance of having two contrasting sets of facts from the same predicates. Despite the fact that the majority of triples of P495 within the Asian part of DLAMA-v1 (Asia-West) has *Japan* as one of the correct labels, a biased model toward predicting *Japan* has a significantly low performance on the opposing set of facts. Consequently, the bias can still be detected.

Regarding the KyKim BERT-base model, language bias toward overpredicting *Korean* is clear for the P103 and the P364 relation predications. The model also shows a bias toward the *Javanese* label for P1412. This bias can be seen in the model's poor performance on the Western part of the benchmark. P19 is a relation predicate on which the model is generally performing well. The most frequent predictions indicate that the model leans toward selecting *Japan* and *United States of America*. However, the model's predictions change according to the underlying culture of the triples and hence demonstrate an ability to memorize facts from both cultures.

Wikidata predicate	T-REx		LAMA		X-FACTR	
	Western countries	Rest of the world	Western countries	Rest of the world	Western countries	Rest of the world
P17 (Country)	321988 (44.0%)	**410512 (56.0%)**	386 (41.6%)	**542 (58.4%)**	457 (45.7%)	**543 (54.3%)**
P19 (Place of birth)	**156579 (68.2%)**	73069 (31.8%)	**730 (77.4%)**	213 (22.6%)	**680 (68.0%)**	320 (32.0%)
P20 (Place of death)	**63250 (76.0%)**	19962 (24.0%)	**734 (77.2%)**	217 (22.8%)	**757 (75.7%)**	243 (24.3%)
P27 (Country of citizenship)	**253402 (63.8%)**	143837 (36.2%)	404 (41.9%)	**560 (58.1%)**	**623 (62.3%)**	377 (37.7%)
P36 (Capital)	4011 (44.8%)	**4936 (55.2%)**	**436 (62.0%)**	267 (38.0%)	418 (41.8%)	**582 (58.2%)**
P39 (Position held)	**7610 (50.1%)**	7581 (49.9%)	380 (42.6%)	**512 (57.4%)**	**504 (50.4%)**	496 (49.6%)
P47 (Shares border with)	**29427 (76.6%)**	9010 (23.4%)	**529 (57.4%)**	393 (42.6%)	**762 (76.2%)**	238 (23.8%)
P101 (Field of work)	**3396 (54.1%)**	2885 (45.9%)	**365 (52.5%)**	330 (47.5%)	**539 (53.9%)**	461 (46.1%)
P103 (Native language)	**6983 (78.4%)**	1926 (21.6%)	**778 (79.7%)**	198 (20.3%)	**787 (78.7%)**	213 (21.3%)
P106 (Occupation)	**203644 (58.7%)**	143177 (41.3%)	**631 (65.9%)**	327 (34.1%)	**578 (57.8%)**	422 (42.2%)
P108 (Employer)	**27119 (91.2%)**	2605 (8.8%)	**371 (96.9%)**	12 (3.1%)	**910 (91.0%)**	90 (9.0%)
P131 (Located in the administrative territorial entity)	**264544 (57.7%)**	194254 (42.3%)	**704 (79.9%)**	177 (20.1%)	**552 (55.2%)**	448 (44.8%)
P136 (Genre)	16396 (17.0%)	**80156 (83.0%)**	**547 (58.8%)**	384 (41.2%)	183 (18.3%)	**817 (81.7%)**
P140 (Religion)	3344 (45.5%)	**4000 (54.5%)**	70 (14.8%)	**403 (85.2%)**	480 (48.0%)	**520 (52.0%)**
P159 (Headquarters location)	**24841 (69.7%)**	10824 (30.3%)	**683 (70.7%)**	283 (29.3%)	**685 (68.5%)**	315 (31.5%)
P190 (Sister city)	**2026 (54.1%)**	1722 (45.9%)	**525 (52.8%)**	470 (47.2%)	**542 (54.2%)**	458 (45.8%)
P276 (Location)	**9239 (65.5%)**	4867 (34.5%)	**554 (57.9%)**	403 (42.1%)	**638 (63.8%)**	362 (36.2%)
P413 (Position played on team / speciality)	**18307 (65.9%)**	9482 (34.1%)	**751 (78.9%)**	201 (21.1%)	**667 (66.7%)**	333 (33.3%)
P463 (Member of)	**13832 (80.0%)**	3452 (20.0%)	112 (49.8%)	**113 (50.2%)**	**807 (80.7%)**	193 (19.3%)
P495 (Country of origin)	**64856 (81.7%)**	14518 (18.3%)	430 (47.4%)	**478 (52.6%)**	**838 (83.8%)**	162 (16.2%)
P530 (Diplomatic relation)	**888 (63.7%)**	505 (36.3%)	**645 (64.8%)**	351 (35.2%)	**633 (63.3%)**	367 (36.7%)
P740 (Location of formation)	**7844 (82.5%)**	1663 (17.5%)	**782 (83.7%)**	152 (16.3%)	**836 (83.6%)**	164 (16.4%)
P937 (Work location)	**6018 (79.7%)**	1535 (20.3%)	**813 (85.2%)**	141 (14.8%)	**801 (80.1%)**	199 (19.9%)
P1001 (Applies to jurisdiction)	**2397 (60.9%)**	1542 (39.1%)	**436 (62.2%)**	265 (37.8%)	**612 (61.2%)**	388 (38.8%)
P1376 (Capital of)	1342 (30.4%)	**3078 (69.6%)**	79 (33.8%)	**155 (66.2%)**	297 (29.7%)	**703 (70.3%)**
P1412 (Languages spoken or published)	**42318 (71.2%)**	17137 (28.8%)	**722 (74.5%)**	247 (25.5%)	**728 (72.8%)**	272 (27.2%)
Total	**1555601 (57.1%)**	1168235 (42.9%)	**13597 (63.6%)**	7794 (36.4%)	**16314 (62.7%)**	9686 (37.3%)

Table A1: The number and percentage of triples belonging to one of the 21 Western countries or to other countries in the T-REx, LAMA, and X-FACTR benchmarks.

Language of Prompt	Model name	$P_{@1}$ *Arab* N=10946	$P_{@1}$ *West* N=13589	$P_{@1}$ *All* N=24535
Arabic	mBERT-base	11.4	**12.8**	12.2
	arBERT	**26.6**	19.3	22.6
English	mBERT-base	19.1	**34.2**	27.5
	BERT-base	24.5	**29.9**	27.5

(a) DLAMA-v1 (Arab-West)

Language of Prompt	Model name	$P_{@1}$ *Asia* N=13479	$P_{@1}$ *West* N=13588	$P_{@1}$ *All* N=27067
Korean	mBERT-base	15.0	**22.6**	18.8
	KyKim	**16.0**	11.8	13.9
English	mBERT-base	27.1	**36.2**	31.7
	BERT-base	**36.4**	30.4	33.4

(b) DLAMA-v1 (Asia-West)

Language of Prompt	Model name	$P_{@1}$ *S.America* N=13071	$P_{@1}$ *West* N=13586	$P_{@1}$ *All* N=26657
Spanish	mBERT-base	22.3	**30.4**	26.4
	BETO	15.5	**25.5**	20.6
English	mBERT-base	24.1	**34.7**	29.5
	BERT-base	24.4	**29.9**	27.2

(c) DLAMA-v1 (South America-West)

Table C2: Performance of mBERT, and monolingual Arabic (arBERT), Korean (KyKim), Spanish (BETO), and English (BERT-base) language models on the three sets of facts of DLAMA-v1 without augmenting the set of objects (i.e.: without applying Step #5).

Contrasting Spanish BETO to English BERT-base: While similar patterns can be found in Tables E5, and E7, a new subtle bias is that BERT-base predicts `Madrid` for more than 50% of the South American triples in P19 (Place of Birth), and P20 (Place of Death) predicates. This might be attributed to the fact that South American names are hard to distinguish from Spanish ones.

F Details of DLAMA

Wikipedia sites: For the Arab, Asian, South American, and Western cultures, representative countries from each region are used as a proxy. Table F8 enumerates the countries representing these cultures and their relevant respective Wikipedia sites.

Probing templates: To probe the models' factual knowledge, natural language templates are used to transform the triples into prompts. The template has two fields for the subject $[X]$ and the object $[Y]$ of the triples. For each triple, the subject fills the subject field while the object field is masked. Models are then fed the prompts and asked to fill in the masked token (i.e., the object). While the templates can affect the predictions of the models, we used the same ones of mLAMA listed in Table F9 to control for the impact that changing the templates might have on the results. In addition to that, we mapped the templates into questions as shown in Table F10 to evaluate the performance of the GPT3.5 model on a subset of DLAMA-v1 (Arab-West).

Relation	# facts (entropy)		Arabic prompts *Accuracy*		English prompts *Accuracy*	
	Arab	West	Arab	West	Arab	West
P30 (Continent)	22 (1.0)	19 (1.0)	63.6	**89.5***	**100.0***	89.5
P36 (Capital)	22 (4.5)	19 (4.2)	**81.8***	63.2	**95.5***	94.7
P37 (Official language)	22 (0.0)	19 (2.5)	**100.0***	89.5	**100.0***	**100.0***
P47 (Shares border with)	22 (2.5)	19 (2.7)	**100.0***	**100.0***	**95.5***	89.5
P190 (Sister city)	67 (4.9)	468 (7.3)	**6.0***	5.6	3.0	**33.1***
P530 (Diplomatic relation)	22 (0.0)	19 (0.0)	63.6	**68.4***	50.0	**84.2***
P1376 (Capital of)	24 (4.3)	26 (4.0)	87.5	**88.5***	**100.0***	92.3

Table D3: The accuracy of the GPT3.5-turbo model for some predicates of the DLAMA-v1 (Arab-West) set.

Relation	# facts (entropy)		Korean prompts $P_{@1}$		English prompts $P_{@1}$	
	Asia	West	Asia	West	Asia	West
P17 (Country)	1000 (2.2)	1000 (2.8)	37.8	**42.1**	**67.1**	45.3
P19 (Place of birth)	1000 (1.7)	1000 (2.7)	**63.1**	55.8	**24.3**	11.9
P20 (Place of death)	1000 (2.6)	1000 (2.8)	23.0	**45.8**	**40.4**	20.7
P27 (Country of citizenship)	1000 (1.5)	1000 (2.4)	**74.0**	53.5	**71.8**	19.5
P30 (Continent)	13 (0.0)	19 (1.0)	**76.9**	31.6	**100.0**	84.2
P36 (Capital)	13 (3.7)	19 (4.2)	**30.8**	21.1	69.2	**84.2**
P37 (Official language)	13 (2.7)	19 (2.5)	**30.8**	26.3	84.6	**100.0**
P47 (Shares border with)	13 (1.7)	19 (2.7)	**0.0**	0.0	76.9	**78.9**
P103 (Native language)	1000 (1.6)	1000 (1.7)	**33.3**	2.3	**84.7**	75.6
P106 (Occupation)	1000 (0.9)	1000 (1.0)	**17.0**	9.4	1.4	**15.9**
P136 (Genre)	1000 (1.0)	1000 (2.5)	0.2	**0.5**	0.8	**6.3**
P190 (Sister city)	387 (7.4)	467 (7.3)	0.0	**1.9**	0.3	**2.8**
P264 (Record label)	1000 (5.3)	1000 (4.8)	**0.3**	0.1	3.3	**6.6**
P364 (Original language of work)	1000 (0.7)	1000 (0.3)	10.5	**18.5**	37.7	**89.1**
P449 (Original network)	1000 (4.6)	1000 (5.0)	**5.1**	0.2	1.1	**10.7**
P495 (Country of origin)	1000 (0.5)	1000 (1.3)	**29.1**	19.2	**79.7**	4.3
P530 (Diplomatic relation)	13 (0.0)	19 (0.0)	**7.7**	5.3	46.2	**68.4**
P1303 (Instrument)	1000 (0.5)	1000 (1.1)	0.4	**1.1**	9.0	**29.5**
P1376 (Capital of)	27 (3.0)	26 (4.0)	**51.9**	26.9	**88.9**	76.9
P1412 (Languages spoken or published)	1000 (1.3)	1000 (1.4)	1.0	**13.4**	**87.4**	86.8
Aggregated statistics	13479 (2.1)	13588 (2.6)	**22.1**	19.5	**38.3**	31.9

Table E4: Detailed P@1 scores of KyKim (Korean prompts) and cased BERT-base (English prompts) on the DLAMA-v1 (Asia-West) set.

Relation	# facts (entropy)		Spanish prompts $P_{@1}$		English prompts $P_{@1}$	
	S.America	West	S.America	West	S.America	West
P17 (Country)	1000 (2.8)	1000 (2.9)	**57.5**	47.7	**63.0**	49.9
P19 (Place of birth)	1000 (2.6)	1000 (2.5)	**2.0**	0.9	**14.6**	8.3
P20 (Place of death)	1000 (2.8)	1000 (2.4)	0.1	**0.6**	0.5	**10.3**
P27 (Country of citizenship)	1000 (2.5)	1000 (2.4)	**19.5**	4.2	**28.9**	14.5
P30 (Continent)	12 (0.0)	19 (1.0)	**91.7**	73.7	**100.0**	73.7
P36 (Capital)	12 (3.6)	19 (4.2)	**83.3**	68.4	66.7	**84.2**
P37 (Official language)	12 (1.2)	19 (2.5)	75.0	**84.2**	75.0	**100.0**
P47 (Shares border with)	12 (1.0)	19 (2.7)	**83.3**	68.4	**91.7**	78.9
P103 (Native language)	1000 (1.1)	1000 (1.8)	34.4	**78.6**	58.5	**74.5**
P106 (Occupation)	1000 (2.1)	1000 (2.5)	6.8	**7.8**	8.3	**12.0**
P136 (Genre)	1000 (2.6)	1000 (2.4)	0.3	**1.7**	2.4	**5.5**
P190 (Sister city)	144 (6.1)	465 (7.4)	**4.9**	1.7	**3.5**	3.0
P264 (Record label)	854 (6.1)	1000 (6.0)	0.0	**0.1**	1.5	**5.6**
P364 (Original language of work)	1000 (1.1)	1000 (0.6)	48.5	**85.1**	60.5	**89.5**
P449 (Original network)	1000 (4.6)	1000 (4.7)	0.3	**0.7**	0.4	**18.7**
P495 (Country of origin)	1000 (2.4)	1000 (1.8)	6.3	**60.0**	**27.3**	10.3
P530 (Diplomatic relation)	12 (0.0)	19 (0.0)	66.7	**68.4**	58.3	**68.4**
P1303 (Instrument)	1000 (1.2)	1000 (1.3)	6.7	**11.7**	17.0	**26.4**
P1376 (Capital of)	13 (3.4)	26 (4.0)	**84.6**	73.1	**84.6**	76.9
P1412 (Languages spoken or published)	1000 (1.2)	1000 (1.7)	20.2	**51.6**	62.9	**89.2**
Aggregated statistics	13071 (2.4)	13586 (2.7)	16.0	**26.5**	26.9	**31.3**

Table E5: Detailed P@1 scores of cased BETO (Spanish prompts) and cased BERT-base (English prompts) on the DLAMA-v1 (South America-West) set.

Relation predicate	Common correct predictions (% of predictions)	Common wrong predictions (% of all predictions)
Probing KyKim with Korean prompts populated with Asian facts		
P17: [X] is located in [Y] .	Japan (30.3%) South Korea (3.0%) Thailand (0.9%)	China (13.8%) United States of America (13.0%) Tonga (9.7%)
P19: [X] was born in [Y] .	Japan (60.1%) South Korea (1.9%) South Chungcheong Province (0.4%)	United States of America (12.8%) South Chungcheong Province (4.0%) South Jeolla (3.2%)
P20: [X] died in [Y] .	Japan (19.2%) Tokyo (2.7%) Gyeonggi Province (0.3%)	United States of America (19.8%) Gyeonggi Province (14.5%) Germany (4.7%)
P27: [X] is [Y] citizen .	Japan (70.2%) South Korea (3.2%) Singapore (0.2%)	Korea (13.5%) South Korea (5.0%) China (3.0%)
P495: [X] was created in [Y] .	Japan (26.2%) South Korea (2.9%)	Jordan (29.7%) South Korea (28.0%) United States of America (6.4%)
P103: The native language of [X] is [Y] .	Korean (31.5%) Japanese (1.5%) Chinese (0.2%)	Korean (66.5%) Hakka (0.1%) Chinese (0.1%)
P364: The original language of [X] is [Y] .	Korean (6.0%) Japanese (4.3%) Chinese (0.1%)	Korean (79.9%) English (8.1%) German (0.4%)
P1412: [X] used to communicate in [Y] .	Vietnamese (0.4%) Javanese (0.3%) Japanese (0.1%)	Javanese (77.2%) Tamil (4.3%) Wu Chinese (3.9%)
Probing KyKim with Korean prompts populated with Western facts		
P17: [X] is located in [Y] .	United States of America (28.2%) France (4.9%) Germany (4.0%)	United States of America (27.4%) Korea (7.6%) China (5.7%)
P19: [X] was born in [Y] .	United States of America (42.0%) France (5.6%) Italy (4.4%)	United States of America (27.0%) Italy (7.7%) Germany (6.8%)
P20: [X] died in [Y] .	United States of America (29.2%) Germany (7.0%) France (5.4%)	Germany (22.2%) United States of America (17.7%) Italy (3.2%)
P27: [X] is [Y] citizen .	United States of America (37.7%) France (11.8%) Italy (1.8%)	United States of America (15.3%) France (10.5%) Korea (9.3%)
P495: [X] was created in [Y] .	United States of America (17.9%) Germany (0.4%) Japan (0.4%)	South Korea (36.4%) Jordan (21.4%) Japan (12.8%)
P103: The native language of [X] is [Y] .	English (2.0%) French (0.3%)	Korean (97.0%) English (0.4%) Japanese (0.2%)
P364: The original language of [X] is [Y] .	English (18.0%) Korean (0.3%) French (0.1%)	Korean (79.3%) English (0.8%) French (0.7%)
P1412: [X] used to communicate in [Y] .	French (7.4%) German (4.9%) Spanish (0.5%)	Javanese (51.2%) German (9.3%) Burmese (9.0%)

Relation predicate	Common correct predictions (% of predictions)	Common wrong predictions (% of all predictions)
Probing BERT-base with English prompts populated with Asian facts		
P17: [X] is located in [Y] .	Japan (48.9%) Thailand (3.4%) Taiwan (3.2%)	China (7.5%) Moscow (5.6%) Taiwan (3.2%)
P19: [X] was born in [Y] .	Tokyo (17.8%) Seoul (2.4%) Vietnam (1.0%)	Tokyo (52.3%) Seoul (7.0%) Beijing (4.1%)
P20: [X] died in [Y] .	Tokyo (26.3%) Beijing (5.8%) Seoul (4.8%)	Beijing (21.0%) Tokyo (16.6%) Paris (7.8%)
P27: [X] is [Y] citizen .	Japan (67.4%) Taiwan (2.4%) Vietnam (1.0%)	Taiwan (12.8%) Singapore (5.4%) Korea (4.6%)
P495: [X] was created in [Y] .	Japan (79.5%) Vietnam (0.1%) Thailand (0.1%)	Japan (5.3%) India (3.5%) Germany (3.0%)
P103: The native language of [X] is [Y] .	Japanese (52.4%) Korean (26.4%) Chinese (3.8%)	English (4.2%) Spanish (1.6%) Wu Chinese (1.5%)
P364: The original language of [X] is [Y] .	Japanese (34.6%) English (2.3%) Chinese (0.3%)	English (50.2%) French (1.8%) Latin (1.5%)
P1412: [X] used to communicate in [Y] .	Japanese (73.5%) Korean (6.8%) Chinese (2.4%)	English (5.6%) Cantonese (2.2%) Japanese (1.5%)
Probing BERT-base with English prompts populated with Western facts		
P17: [X] is located in [Y] .	France (15.5%) Germany (8.9%) Spain (8.3%)	Georgia (12.6%) Moscow (7.7%) Canada (6.3%)
P19: [X] was born in [Y] .	Paris (4.5%) London (1.2%) Rome (1.0%)	Chicago (24.2%) London (21.4%) Paris (10.5%)
P20: [X] died in [Y] .	Paris (11.2%) Rome (3.7%) London (2.5%)	Paris (30.7%) London (19.9%) Rome (9.4%)
P27: [X] is [Y] citizen .	France (11.8%) Italy (3.4%) Canada (1.0%)	British America (35.8%) Singapore (19.7%) Austria (4.9%)
P495: [X] was created in [Y] .	France (1.5%) Germany (1.0%) Japan (0.5%)	Japan (60.2%) England (10.7%) Germany (5.0%)
P103: The native language of [X] is [Y] .	English (53.1%) French (12.6%) German (3.4%)	Spanish (6.9%) French (3.9%) German (3.7%)
P364: The original language of [X] is [Y] .	English (87.0%) French (0.7%) German (0.5%)	Latin (2.3%) English (1.9%) French (1.7%)
P1412: [X] used to communicate in [Y] .	English (61.7%) French (12.1%) German (4.3%)	English (4.6%) Spanish (2.4%) Arabic (1.4%)

Table E6: The most common predictions for monolingual Korean and English BERT models when probed by DLAMA-v1 (Asia-West) with English and Korean prompts, respectively. Purple culturally related prediction, Blue bell culturally proximate prediction, Light Orange culturally proximate prediction to another culture, Orange culturally related prediction to the other culture. **Note:** The Korean prompts/entities are translated for clarity.

Relation predicate	Common correct predictions (% of predictions)	Common wrong predictions (% of all predictions)
Probing BETO with Spanish prompts populated with South America facts		
P17: [X] is located in [Y] .	Brazil (17.6%) Argentina (14.9%) Chile (7.9%)	Mexico (12.0%) Curaçao (8.1%) Venezuela (4.3%)
P19: [X] was born in [Y] .	Buenos Aires (1.5%) Lima (0.2%) Brazil (0.1%)	Altötting (91.0%) Buenos Aires (5.6%) Madrid (0.3%)
P20: [X] died in [Y] .	Aripuanã (0.1%)	Aripuanã (99.6%) Buenos Aires (0.1%) Caracas (0.1%)
P27: [X] is [Y] citizen .	Brazil (13.0%) Colombia (4.3%) Chile (1.4%)	Colombia (39.7%) Taiwan (9.0%) Mexico (5.5%)
P495: [X] was created in [Y] .	Argentina (1.3%) Chile (1.2%) Brazil (1.1%)	United States of America (29.6%) Río de la Plata (23.4%) Kingdom of Portugal (16.6%)
P103: The native language of [X] is [Y] .	Spanish (17.6%) Portuguese (15.4%) English (1.4%)	English (48.2%) Spanish (14.1%) French (2.2%)
P364: The original language of [X] is [Y] .	Spanish (39.1%) Portuguese (7.5%) English (1.9%)	English (36.1%) Spanish (13.7%) French (0.5%)
P1412: [X] used to communicate in [Y] .	Spanish (13.4%) English (6.5%) Portuguese (0.2%)	English (70.6%) Spanish (5.8%) Latin (2.6%)
Probing BETO with Spanish prompts populated with Western facts		
P17: [X] is located in [Y] .	France (13.1%) Spain (10.6%) United States of America (10.2%)	Mexico (15.5%) United States of America (5.7%) Canada (3.2%)
P19: [X] was born in [Y] .	Paris (0.5%) Rome (0.2%) Altötting (0.1%)	Altötting (95.6%) Paris (2.4%) Rome (0.4%)
P20: [X] died in [Y] .	Paris (0.3%) Rome (0.2%) Madrid (0.1%)	Aripuanã (98.8%) Paris (0.2%) Rome (0.1%)
P27: [X] is [Y] citizen .	France (1.4%) Italy (1.1%) Spain (0.5%)	Taiwan (25.3%) Australia (21.7%) Socialist Republic of Romania (7.3%)
P495: [X] was created in [Y] .	United States of America (54.6%) France (2.6%) Spain (2.3%)	United States of America (12.5%) Kingdom of Portugal (5.1%) Río de la Plata (4.9%)
P103: The native language of [X] is [Y] .	English (63.3%) French (10.7%) German (1.7%)	English (18.4%) French (1.2%) Spanish (1.2%)
P364: The original language of [X] is [Y] .	English (80.4%) Spanish (2.8%) Italian (0.7%)	Spanish (10.5%) English (3.5%) French (0.3%)
P1412: [X] used to communicate in [Y] .	English (41.3%) French (6.0%) German (2.9%)	English (43.6%) Spanish (3.1%) Latin (1.3%)
Relation predicate	**Common correct predictions (% of predictions)**	**Common wrong predictions (% of all predictions)**
Probing BERT-base with English prompts populated with South American facts		
P17: [X] is located in [Y] .	Brazil (18.1%) Argentina (16.3%) Chile (8.1%)	Bolivia (6.5%) Mexico (5.1%) Spain (3.5%)
P19: [X] was born in [Y] .	Brazil (14.0%) Argentina (0.3%) Bolivia (0.1%)	Madrid (54.1%) Rome (6.4%) Milan (3.9%)
P20: [X] died in [Y] .	Peru (0.2%) Brazil (0.2%) London (0.1%)	Madrid (55.1%) Paris (16.8%) Rome (11.4%)
P27: [X] is [Y] citizen .	Brazil (18.0%) Argentina (9.9%) Italy (0.3%)	Mexico (25.0%) Argentina (15.8%) Honduras (6.0%)
P495: [X] was created in [Y] .	Brazil (19.7%) Argentina (2.2%) Chile (1.8%)	Mexico (24.4%) Japan (11.3%) Spain (8.7%)
P103: The native language of [X] is [Y] .	Portuguese (34.2%) Spanish (23.2%) English (0.6%)	Spanish (20.5%) Italian (4.8%) English (3.6%)
P364: The original language of [X] is [Y] .	Spanish (40.7%) Portuguese (17.1%) English (2.6%)	English (19.4%) Spanish (6.6%) Latin (5.5%)
P1412: [X] used to communicate in [Y] .	Spanish (44.7%) Portuguese (14.5%) English (2.6%)	Spanish (16.4%) English (12.0%) Italian (3.0%)
Probing BERT-base with English prompts populated with Western facts		
P17: [X] is located in [Y] .	France (16.9%) Spain (12.7%) Germany (7.9%)	Georgia (9.5%) Canada (5.1%) Lebanon (2.9%)
P19: [X] was born in [Y] .	Berlin (2.3%) Paris (1.7%) London (1.7%)	Berlin (36.7%) Chicago (13.0%) London (12.9%)
P20: [X] died in [Y] .	Paris (5.1%) London (1.8%) Rome (1.1%)	Munich (23.0%) Paris (22.8%) Berlin (15.0%)
P27: [X] is [Y] citizen .	France (8.6%) Austria (2.0%) Italy (1.9%)	Austria (36.8%) British America (26.1%) Netherlands (4.2%)
P495: [X] was created in [Y] .	France (4.8%) Spain (1.6%) Germany (1.5%)	Japan (53.1%) England (10.3%) Germany (5.3%)
P103: The native language of [X] is [Y] .	English (48.8%) French (15.1%) German (3.4%)	Spanish (7.4%) German (3.7%) French (3.0%)
P364: The original language of [X] is [Y] .	English (83.8%) Spanish (2.9%) German (1.3%)	English (2.3%) Latin (2.0%) French (1.8%)
P1412: [X] used to communicate in [Y] .	English (38.8%) German (34.0%) French (10.1%)	English (5.4%) Spanish (1.3%) German (0.8%)

Table E7: The most common predictions for monolingual Spanish and English BERT models when probed by DLAMA-v1 (South America-West) with English and Spanish prompts, respectively. Purple culturally related prediction, Blue bell culturally proximate prediction, Light Orange culturally proximate prediction to another culture, Orange culturally related prediction to the other culture. **Note:** The Spanish prompts/entities are translated for clarity.

Cultures	Country	Wikipedia sites used for articles
Arab Cultures	22 countries of the Arab League	Arabic (ar), English (en), French (fr)
Western Cultures	Australia	English (en)
	Canada	English (en), French (fr)
	New Zealand	English (en), Mori (mi)
	USA	English (en)
	Andorra	Catalan (ca), English (en)
	Italy	Italian (it), English (en)
	Liechtenstein	German (de), English (en)
	Monaco	French (fr), English (en)
	Portugal	Portuguese (pt), English (en)
	San Marino	Italian (it), English (en)
	Spain	Spanish (es), English (en)
	Austria	German (de), English (en)
	Belgium	German (de), French (fr), Dutch (nl), English (en)
	France	French (fr), English (en)
	Germany	German (de), English (en)
	Ireland	Irish (ga), English (en)
	Luxembourg	Luxembourgish (lb), French (fr), German (de), English (en)
	Netherlands	Dutch (nl), English (en)
	Switzerland	German (de), French (fr), Italian (it), Romansh (rm), English (en)
	UK	English (en)
Asian Cultures	China	English (en), Chinese (zh)
	Indonesia	English (en), Indonesian (id)
	Japan	English (en), Japanese (ja)
	Malaysia	English (en), Malay (ms)
	Mongolia	English (en), Chinese (zh)
	Myanmar	English (en), Burmese (my)
	North Korea	English (en), Korean (ko)
	Philippines	English (en)
	Singapore	English (en), Malay (ms)
	South Korea	English (en), Korean (ko)
	Taiwan	English (en), Chinese (zh)
	Thailand	English (en), Thai (th)
	Vietnam	English (en), Vietnamese (vi)
South American Cultures	Argentina	English (en), Spanish (es)
	Bolivia	English (en), Spanish (es)
	Brazil	English (en), Portugese (pt)
	Chile	English (en), Spanish (es)
	Colombia	English (en), Spanish (es)
	Ecuador	English (en), Spanish (es)
	Guyana	English (en)
	Paraguay	English (en), Spanish (es)
	Peru	English (en), Spanish (es)
	Suriname	Dutch (nl), English (en)
	Uruguay	English (en), Spanish (es)
	Venezuela	English (en), Spanish (es)

Table F8: The list of Countries and their respective Wikipedia sites used for representing the four different cultures. The English Wikipedia is used for all the countries.

Predicate	English template	Arabic template	Korean template	Spanish template
P17 (Country)	[X] is located in [Y] .	يقع [X] في [Y].	[X]는 [Y]에 있습니다.	[X] se encuentra en [Y].
P19 (Place of birth)	[X] was born in [Y] .	ولد [X] في [Y].	[X]는 [Y]에서 태어났습니다.	[X] nació en [Y].
P20 (Place of death)	[X] died in [Y] .	توفي [X] في [Y].	[X]는 [Y]에서 사망했습니다.	[X] murió en [Y].
P27 (Country of citizenship)	[X] is [Y] citizen .	[X] مواطن [Y].	[X]는 [Y] 시민입니다.	[X] es [Y] ciudadano.
P30 (Continent)	[X] is located in [Y] .	يقع [X] في [Y].	[X]는 [Y]에 있습니다.	[X] se encuentra en [Y].
P36 (Capital)	The capital of [X] is [Y] .	عاصمة [X] هي [Y].	[X]의 수도는 [Y]입니다.	La capital de [X] es [Y].
P37 (Official language)	The official language of [X] is [Y] .	اللغة الرسمية لـ [X] هي [Y].	[X]의 공식 언어는 [Y]입니다.	El idioma oficial de [X] es [Y].
P47 (Shares border with)	[X] shares border with [Y] .	[X] تشترك في الحدود مع [Y].	[X]는 [Y]와 (과) 국경을 공유합니다.	[X] comparte frontera con [Y].
P103 (Native language)	The native language of [X] is [Y] .	اللغة الأصلية لـ [X] هي [Y].	[X]의 모국어는 [Y]입니다.	El idioma nativo de [X] es [Y].
P106 (Occupation)	[X] is a [Y] by profession .	[X] هي [Y] حسب المهنة.	[X]는 직업 별 [Y]입니다.	[X] es una [Y] de profesión.
P136 (Genre)	[X] plays [Y] music .	[X] يعزف موسيقى [Y].	[X]는 [Y] 음악을 재생합니다.	[X] reproduce música [Y].
P190 (Sister city)	[X] and [Y] are twin cities .	[X] و [Y] مدينتان توأمتان.	[X]와 [Y]는 쌍둥이 도시입니다.	[X] e [Y] son ciudades gemelas.
P264 (Record label)	[X] is represented by music label [Y] .	[X] يمثلها العلامة الموسيقية [Y].	[X]는 음악 레이블 [Y]로 표시됩니다.	[X] está representado por el sello musical [Y].
P364 (Original language of work)	The original language of [X] is [Y] .	اللغة الأصلية لـ [X] هي [Y].	[X]의 원래 언어는 [Y]입니다.	El idioma original de [X] es [Y].
P449 (Original network)	[X] was originally aired on [Y] .	تم بث [X] في الأصل على [Y].	[X]는 원래 [Y]에 방영되었습니다.	[X] se emitió originalmente en [Y].
P495 (Country of origin)	[X] was created in [Y] .	تم إنشاء [X] في [Y].	[X]는 [Y]에 작성되었습니다.	[X] se creó en [Y].
P530 (Diplomatic relation)	[X] maintains diplomatic relations with [Y] .	[X] تقيم علاقات دبلوماسية مع [Y].	[X]는 [Y]와의 외교 관계를 유지합니다.	[X] mantiene relaciones diplomáticas con [Y].
P1303 (Instrument)	[X] plays [Y] .	[X] يلعب [Y].	[X]는 [Y]를 재생합니다.	[X] reproduce [Y].
P1376 (Capital of)	[X] is the capital of [Y] .	[X] هي عاصمة [Y].	[X]는 [Y]의 수도입니다.	[X] es la capital de [Y].
P1412 (Languages spoken or published)	[X] used to communicate in [Y] .	يستخدم [X] للتواصل في [Y].	[X]는 [Y]에서 통신하는 데 사용됩니다.	[X] solía comunicarse en [Y].

Table F9: mLAMA's templates that are also adapted in DLAMA.

Predicate	English Question	Arabic Question
P30 (Continent)	Where is "[X]" located in? Reply with a name of a continent only.	أين يقع "[X]"؟ أجب باسم قارة فقط
P36 (Capital)	What is the capital of "[X]"? Reply with the name of the city only.	ما هي عاصمة "[X]"؟ أجب باسم المدينة فقط
P37 (Official Language)	What is the official language of "[X]"? Reply with the language name only.	ما هي اللغة الرسمية ل "[X]"؟ أجب باسم لغة فقط
P47 (Shares border with)	What is the country that shares border with "[X]"? Reply with a country name only.	ما هي الدولة التي تشترك حدودها مع "[X]"؟ أجب باسم دولة فقط
P190 (Sister city)	What is the twin city of "[X]"? Reply with the name of the city only.	ما هي المدينة التوأم لمدينة "[X]"؟ أجب باسم المدينة فقط
P530 (Diplomatic realtion)	What is the country that maintains dimplomatic relations with "[X]"? Reply with a country name only.	ما هي الدولة التي تقيم علاقات دبلوماسية مع "[X]"؟ أجب باسم دولة فقط
P1376 (Capital of)	What is the country of which the capital is "[X]"? Reply with a country name only.	ما هي الدولة التي عاصمتها "[X]"؟ أجب باسم دولة فقط

Table F10: The mapping of six of mLAMA's templates to questions that can be used to evaluate the GPT3.5-turbo model.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☑ A2. Did you discuss any potential risks of your work?
Ethical Considerations section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 2, 3, 4

☑ B1. Did you cite the creators of artifacts you used?
Sections 2, 3, 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The benchmarks are dumps of factual triples from Wikidata.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 4, Appendix - Section E

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 4, Appendix - Section E

C ☑ Did you run computational experiments?

Section 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
The codebase is linked to in the Introduction section

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.